Published in final edited form as:

*Biometrika*. 2014 September ; 101(3): 625–640. doi:10.1093/biomet/asu017.

# Multicategory angle-based large-margin classification

**Chong Zhang** and **Yufeng Liu**

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

Chong Zhang: chongz@live.unc.edu; Yufeng Liu: yfliu@email.unc.edu

## Summary

Large-margin classifiers are popular methods for classification. Among existing simultaneous multicategory large-margin classifiers, a common approach is to learn $k$ different functions for a $k$-class problem with a sum-to-zero constraint. Such a formulation can be inefficient. We propose a new multicategory angle-based large-margin classification framework. The proposed angle-based classifiers consider a simplex-based prediction rule without the sum-to-zero constraint, and enjoy more efficient computation. Many binary large-margin classifiers can be naturally generalized for multicategory problems through the angle-based framework. Theoretical and numerical studies demonstrate the usefulness of the angle-based methods.

### Some key words

Hard classification; Probability estimation; Soft classification; Support vector machine

## 1. Introduction

Classification is an important supervised learning technique with numerous applications. Given a training dataset with subjects having both covariates and class labels, the goal is to build a classifier that predicts the label for a new subject. Existing techniques for classification include Fisher linear discriminant analysis, logistic regression, support vector machines and boosting; see Hastie et al. (2009).

Large-margin classifiers in machine learning have attracted a lot of attention in recent years. Typically a large-margin classification method can be written in the regularization framework of loss plus penalty. The loss term measures the goodness of fit of the resulting classifier, and the penalty term controls model complexity and helps to prevent overfitting. Many existing binary classification methods belong to the large-margin classification framework, such as adaboost (Freund & Schapire, 1997), penalized logistic regression (Lin et al., 2000), import vector machines (Zhu & Hastie, 2005), support vector machines (Boser et al., 1992), ψ-learning (Shen et al., 2003), and large-margin unified machines (Liu et al., 2011).

Supplementary material

Supplementary material available at *Biometrika* online includes the large-margin unified loss function, proofs of the theorems, more simulation examples and results, details for the Glioblastoma Multiforme Cancer data, and a summary of attributes of the real datasets.

Although binary large-margin classifiers are popular, extensions are needed for multicategory problems. One simple approach is to deal with a multicategory classification problem via a sequence of binary classifiers. Both the one-versus-one and one-versus-rest approaches are sequential binary methods, which are invariant with respect to the order of the class labels. These methods can be suboptimal in certain situations. For example, the one-versus-rest support vector machine method can be inconsistent when there is no dominating class (Lee et al., 2004; Liu & Yuan, 2011). Hence, it is desirable to study frameworks that consider all classes simultaneously for multicategory classification problems.

For a simultaneous classification problem with $k > 2$ classes, a common approach in the literature is to map the covariates to a classification function vector with length $k$. In this framework, one associates each class with a coordinate in the function vector. The prediction rule assigns the class label to the category that corresponds to the maximum element within the function vector. A sum-to-zero constraint on the function vector is commonly used to reduce the parameter space as well as to obtain desirable theoretical properties. Many proposed methods follow this framework. See, for example, Vapnik (1998), Crammer & Singer (2001), Lee et al. (2004), Zhu & Hastie (2005), Liu & Shen (2006), Zhu et al. (2009) and Liu & Yuan (2011). However, as binary large-margin classifiers estimate a single function to perform classification using its sign, for a $k$-class problem, it should be sufficient to consider a $(k - 1)$-dimensional classification function space. Therefore, existing simultaneous classifiers with the sum-to-zero constraint can be inefficient. The corresponding computation is more involved as well. Apart from inefficiency, the geometric interpretation of maximum separation for binary support vector machines is well understood, while the extension to multicategory support vector machines is much less clear.

To overcome these difficulties, we propose a new multicategory angle-based large-margin classification technique that not only provides a natural generalization of binary large-margin methods, but also overcomes the disadvantages of the existing methods mentioned above. In particular, we consider a $k$-vertex simplex structure in an Euclidean space, whose dimension is $k - 1$, one less than the number of categories (Hill & Doucet, 2007; Lange & Wu, 2008). Our notation of a $k$-vertex simplex is centred at the origin with equal pairwise distances among the vertices. The classification function vector maps the covariates of a given instance to the Euclidean space wherein the simplex lies. In the $(k - 1)$-dimensional space, each vertex of the simplex represents one class and is a $(k - 1)$-dimensional vector. The prediction rule assigns the given instance to the category whose corresponding vertex vector has the smallest angle with respect to the mapped classification function vector. The details of class label representations using a simplex and the prediction rule are introduced in § 2.

Compared to the regular simultaneous multicategory classification methods, the angle-based method has several attractive properties. First, the geometric interpretation of the least angle prediction rule is very easy to understand through our newly defined classification regions. See Fig. 1(a) in § 2. The corresponding functional margin can be directly generalized from the binary to the multicategory case. See Chapter 6 of Cristianini & Shawe-Taylor (2000)

for the definition of functional margin. The minimizers of the new multicategory large-margin unified machines using the angle-based structure help us to better understand the transition behaviours from soft to hard classifiers (Wahba, 2002; Liu et al., 2011). Second, our angle-based method enjoys a lower computational cost than the regular procedure with the sum-to-zero constraint, as we implicitly transfer the constraint onto our newly defined functional margins. Furthermore, we show that the angle-based method has many desirable theoretical properties. In particular, we provide some theoretical insights on the advantages of our angle-based method in § 3·5. Despite the sum-to-zero constraint, using $k$ classification functions instead of $k - 1$ can introduce extra variability in the estimated classifier. This can reduce the classification performance of the regular simultaneous classifiers. In linear learning problems with high dimensional predictors, this extra variability can be large. Consequently, the angle-based methods can significantly outperform the regular simultaneous methods. We confirm this insight numerically by comparing our angle-based methods with some existing simultaneous multicategory classifiers. We also show that the numerical performance of our angle-based method is competitive for nonlinear problems using kernel learning.

## 2. Methodology

### 2·1. Binary large-margin classifiers

Suppose we are given a training dataset, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, obtained from an unknown underlying distribution $P(x, y)$. In classification problems, one important goal is to find a classifier so that the corresponding misclassification rate is minimized. Our focus in this paper is on large-margin classifiers. We first review binary large-margin classifiers in this section, and then introduce the new angle-based methodology for multicategory problems in § 2·2.

For simplicity, we first discuss binary problems, with $y \in \{1, -1\}$. In that case, we have a single classification function $f$ with sign ($f$) as the corresponding prediction rule. Specifically, for an instance with covariates $x$, the predicted class is $\hat{y} = 1$ if $f(x) \geq 0$, and $\hat{y} = -1$ otherwise. Correct classification occurs if and only if $yf(x) > 0$. This quantity $yf(x)$ is known as the functional margin. Given the classification function $f$, the theoretical 0–1 loss can be expressed as $L\{yf(x)\} = I[y \neq \text{sign}\{f(x)\}]$, where $I(\cdot)$ is the indicator function. Because $L$ is discontinuous and nonconvex, direct minimization of the total empirical 0–1 loss is difficult. To overcome this challenge, a common approach is to use a surrogate convex loss function in place of $L$. Let $\ell(\cdot)$ be a convex upper bound of $L$. A large-margin classifier solves the minimization problem $\min_{f \in F} n^{-1} \sum_{i=1}^{n} \ell\{yf(x_i)\} + \lambda J(f)$, or alternatively $\min_{f \in F} n^{-1} \sum_{i=1}^{n} \ell\{yf(x_i)\}$, subject to $J(f) \leq s$, where $F$ is the functional space of interest, $J(f)$ is the penalty term on $f$ to control its complexity and consequently prevent the resulting classifier from overfitting, and $\lambda$, $s$ are tuning parameters that balance the goodness of fit and the penalty terms.

The use of the loss $\ell$ circumvents the difficulty of minimizing the 0–1 loss $L$. In the literature, many binary classification methods use a nonincreasing $\ell$ to encourage a large functional margin $yf(x)$. For example, the support vector machine uses the hinge loss $\ell(yf) =$

$(1 - yf)_+$ (Vapnik, 1998;Wahba, 1999), logistic regression uses the deviance loss $\ell(yf) = \log\{1 + \exp(-yf)\}$ (Lin et al., 2000; Zhu & Hastie, 2005), boosting approximately uses the exponential loss $\ell(yf) = \exp(-yf)$ (Freund & Schapire, 1997), and the large-margin unified machine uses the large-margin unified loss function (Liu et al., 2011). See the Supplementary Material for details. This family provides a convenient platform for studying the transition behaviour from soft to hard binary classifiers (Wahba, 2002). When $c = 0$, the large-margin unified machine loss is a typical soft classifier, and when $c \to \infty$, the large-margin unified machine loss becomes the hinge loss in the standard support vector machine, which is a typical hard classifier.

### 2·2. Multicategory angle-based large-margin classifiers

Multicategory problems are prevalent in practice. In that case, it is desirable to extend binary large-margin classifiers to multicategory versions.

For multicategory problems, let $Y \in \{1, \ldots, k\}$, where $k > 2$ is the number of classes. The regular simultaneous procedure maps $x$ to $f(x) \in \mathbb{R}^k$, and the corresponding prediction rule is $\hat{y} = \text{argmax}_j f_j(x)$, where $f_j$ is the $j$th element of $f$. A sum-to-zero constraint on $f$ is typically imposed, and many statistical properties such as Fisher consistency can be derived. This constraint is commonly used in the literature, see for example, Lee et al. (2004), Wang & Shen (2007), Zhu et al. (2009), Liu & Yuan (2011), and Zhang & Liu (2013). However, since only one function is needed for binary classification, it should be sufficient to use $k - 1$ functions for $k$-class problems. Hence, to construct $k$ functions and remove the redundancy by the sum-to-zero constraint can be inefficient. Furthermore, the corresponding optimization problem can be more challenging as well. See the discussion in § 4 for more details. The angle-based method we introduce next overcomes the difficulties mentioned above. We give some geometric explanations on how angle-based classifiers work for multicategory classification problems. Without the sum-to-zero constraint, the computational speed of our angle-based methods can be significantly faster than the regular methods.

To begin with, we define a specific simplex in a $(k - 1)$-dimensional space, as studied in Lange & Wu (2008). We propose large-margin classifiers using the simplex coding, while the classifiers introduced in Lange & Wu (2008) and Wu & Wu (2012) are regression-based methods. In this setting, a simplex is defined as a $k$-regular polyhedron in a $(k - 1)$-dimensional Euclidean space. For example, when $k$ =3, the simplex is an equilateral triangle in $\mathbb{R}^2$, and when $k = 4$, it is a regular tetrahedron in $\mathbb{R}^3$. For a multicategory classification problem with $k$ classes, define $W$ as the collection of $k$ vectors in $\mathbb{R}^{k-1}$ with elements

$$W_j = \begin{cases} (k-1)^{-1/2}\zeta, & j=1, \\ -(1+k^{1/2})/\{(k-1)^{3/2}\}\zeta + \{k/(k-1)\}^{1/2} e_{j-1}, & 2 \le j \le k, \end{cases}$$

where $\zeta$ is a vector with length $k - 1$ and each element 1, and $e_j$ is a vector in $\mathbb{R}^{k-1}$ such that its every element is 0, except the $j$th element is 1. One can check that $W = \{W_1, \ldots, W_k\}$ forms a simplex with $k$ vertices in the $(k - 1)$-dimensional space. The centre of $W$ is at the origin, and each $W_j$ has Euclidean norm 1.

The set $\{W_1, \ldots, W_k\}$ defines $k$ directions in $\mathbb{R}^{k-1}$, and the angles between any two directions are equal. Every vector in $\mathbb{R}^{k-1}$ generates $k$ different angles with respect to $\{W_1, \ldots, W_k\}$. Assume that the angles are in $[0, \pi]$. Let $W_j$ represent class $j$. For a fitted $\hat{f}$, our method is to map $x$ to $\hat{f}(x) \in \mathbb{R}^{k-1}$, and predict $\hat{y}$ to be the class whose corresponding angle is the smallest. In particular, we let $\hat{y} = \text{argmin}_j \angle(W_j, \hat{f})$, where $\angle(\cdot, \cdot)$ denotes the angle between two vectors. Based on the prediction rule, we can split the space $\mathbb{R}^{k-1}$ into $k$ disjoint classification regions, whose definition is given below, together with a boundary set.

**Definition 1—***A classification region with respect to class j, $C_j$ ($j = 1, \ldots, k$), is the subset of $\mathbb{R}^{k-1}$, such that if $\hat{f} \in C_j$, then $\hat{y} = j$.*

The classification regions are closely connected to the prediction rule. In Fig. 1(a) we plot the classification regions for the angle-based method with $k = 3$. For a given vector $f \in \mathbb{R}^{k-1}$, $\angle(W_j, f)$ is smallest if and only if the projection of $f$ on $W_j$ is the largest for $j = 1, \ldots, k$. In other words, $\angle(W_j, f)$ being smallest is equivalent to $\langle f, W_j \rangle > \langle f, W_l \rangle$, for $l \neq j$, where $\langle x_1, x_2 \rangle = x_1^{\mathrm{T}} x_2$ is the inner product of vectors $x_1$, $x_2$, and $x^{\mathrm{T}}$ is the transpose of $x$. Moreover, the smaller $\angle(W_j, f)$ is, the larger $\langle f, W_j \rangle$ is. Therefore, for any binary large-margin loss function $\ell(\cdot)$, we propose to solve the following optimization for the angle-based classifier

$$\min_{f \in F} n^{-1} \sum_{i=1}^{n} \ell\{\langle f(x_i), W_{y_i} \rangle\} + \lambda J(f). \quad (1)$$

By duality with convex $\ell$ and $J$, (1) is equivalent to the optimization problem

$$\min_{f \in F} n^{-1} \sum_{i=1}^{n} \ell\{\langle f(x_i), W_{y_i} \rangle\}, \quad (2)$$

subject to $J(f) \leq s$. Our prediction rule is now equivalent to $\hat{y} = \text{argmax}_j \langle \hat{f}(x), W_j \rangle$. The value $\langle f(x), W_y \rangle$, which is equivalent to the projected length of $f(x)$ on $W_y$, can be viewed as a generalized functional margin of $(x, y)$, and $\ell\{\langle f(x), W_y \rangle\}$ measures the loss of assigning the label $y$ to $x$, in terms of the inner product between $f(x)$ and $W_y$. From this perspective, the loss function in (1) and (2) encourages $\angle\{W_{y_i}, f(x_i)\}$ to be as small as possible, or equivalently, it encourages the projection of $f(x_i)$ on $W_{y_i}$ to be as large as possible. Notice that $\sum_{j=1}^{k} \langle W_j, f \rangle = 0$, which means we implicitly transfer the sum-to-zero constraint on the classification function vector $f$ in the regular simultaneous multicategory classification to our new functional margins $\langle W_j, f \rangle$. Hence, without an explicit constraint on the classification function $f$, we reduce the computational burden of the optimization problem. As we will see in § § 5 and 6, it is much more efficient to compute the angle-based classification solution than the regular multicategory classifiers with $k$ functions.

The representation of the multicategory class label using the simplex structure was used in the literature previously. Some special cases with certain loss functions have been studied in the literature using the simplex class coding, such as the ε-insensitive loss (Lange & Wu, 2008; Wu & Wu, 2012), boosting (Saberian & Vasconcelos, 2011) and support vector machines (Hill & Doucet, 2007). Our method is more general and covers general large-

margin classifiers. The methods proposed by Saberian & Vasconcelos (2011) and Hill & Doucet (2007) can be viewed as special cases of the angle-based structure.

## 3. Statistical properties

### 3·1. Fisher consistency and theoretical minimizer

Fisher consistency is also known as classification calibration (Bartlett et al., 2006; Tewari & Bartlett, 2007), or infinite sample consistency (Zhang, 2004a). Intuitively, when the functional space $F$ is large enough and we have infinitely many samples, the prediction rule for a Fisher consistent large-margin classifier should yield the minimum misclassification rate. It is a fundamental requirement of a classification loss function.

Define the class conditional probabilities $P_j = \text{pr}(Y = j \mid X = x)$ ($j = 1, \ldots, k$). Under our angle-based prediction rule, Fisher consistency requires that for a given $x$ such that $P_y > P_j$ ($j \neq y$) for some $y \in \{1,\ldots, k\}$, the vector $f^*(x)$ that minimizes $E[\ell\{\langle f(X), W_Y\rangle\} \mid X = x]$ satisfies that $y = \text{argmax}_j \langle f^*(x), W_j\rangle$ and such an argmax is unique. In other words, Fisher consistency requires that $f^* \in C_y$, where $f^*$ is the theoretical minimizer. Next we show that the angle-based method is Fisher consistent if the loss $\ell$ has a negative derivative function.

**Theorem 1—***The angle-based classification loss function $\ell$ in* (1) *and* (2) *is Fisher consistent if the derivative $\ell'$ exists and $\ell'(x) < 0$ for all $x$.*

The logistic loss, the exponential loss, and the large-margin unified machine family with $c < \infty$ meet the conditions of Theorem 1, and thus are all Fisher consistent. For the support vector machine, the hinge loss is not differentiable and hence does not satisfy the condition of Theorem 1. As the large-margin unified machine family includes the hinge loss as a special case with $c \to \infty$, we show that the multicategory angle-based support vector machine is not Fisher consistent, by studying the theoretical minimizer $f^*$ of the entire large-margin unified machine family. The next theorem gives the explicit expression of $f^*$ in terms of $P_j$ ($j = 1, \ldots, k$).

**Theorem 2—***Consider the angle-based method with $\ell$ in the large-margin unified machine family. Assume that $P_1 > \cdots > P_k > 0$. The theoretical minimizer $f^*$ satisfies*

$$\langle W_j, f^*\rangle = \begin{cases} \{(P_j/P_k)^{1/(a+1)}a - a + c\}/(1+c), & j \neq k, \\ -\sum_{u=1}^{k-1}\{(P_u/P_k)^{1/(a+1)}a - a + c\}/(1+c), & j = k, \end{cases}$$

*for $a \in (0, \infty)$ and $c \in [0, \infty]$. Hence*

$f^* = (k-1)\sum_{j=1}^{k-1}\{(P_j/P_k)^{1/(a+1)}a - a + c\}(W_j - W_k)/\{k(1+c)\}$. *For $a = \infty$ and $c \in [0, +\infty]$, we have that*

$$\langle W_j, f^*\rangle = \begin{cases} \{\log(P_j/P_k) + c\}/(1+c), & j \neq k, \\ -\sum_{u=1}^{k-1}\{\log(P_u/P_k) + c\}/(1+c), & j = k. \end{cases}$$

*In this case,* $f^*=(k-1)\sum_{j=1}^{k-1}\{\log(P_j/P_k)+c\}(W_j-W_k)/\{k(1+c)\}$.

For any fixed $(P_1, \ldots, P_k)$ in Theorem 2 and $a < \infty$, we have $\langle W_1, f^*\rangle - \langle W_2, f^*\rangle = a\{(P_1/P_k)^{1/(a+1)} - (P_2/P_k)^{1/(a+1)}\}/(c+1)$. As $c$ increases, the difference in the functional margins $\langle W_1, f^*\rangle - \langle W_2, f^*\rangle$ decreases, and consequently the difference between $\angle(W_1, f^*)$ and $\angle(W_2, f^*)$ decreases. Hence, one can verify that as $c$ increases, the theoretical minimizer $f^*$ stays in $C_1$ and moves closer to the boundary separating different classification regions. When $c \to \infty$ with $\ell$ the hinge loss, the theoretical minimizer $f^*$ satisfies $\langle W_j, f^*\rangle = 1$ ($j = 1, \ldots, k-1$) and $\langle W_k, f^*\rangle = -k+1$. Therefore, $f^*$ is on the boundary set, and consequently the multicategory angle-based support vector machine is not Fisher consistent. See Fig. 1(b) for an illustration. The situation is similar with $a \to \infty$. To overcome this difficulty, we propose to approximate the hinge loss with a large-margin unified machine loss function with large $c$. Under the angle-based structure, we call this large-margin unified machine loss with a large but finite $c$ the approximate support vector machine loss. By Theorem 2, the angle-based support vector machine using the approximate support vector machine loss is Fisher consistent.

In Theorem 2 we assume that $P_1 > \cdots > P_k > 0$ only for simplicity of expression. When $P_i$ and $P_{i+1}$ are the same, the theoretical minimizer is not unique. When $P_k = 0$, $f^*$ is unbounded. We discuss this further in the Supplementary Material.

### 3·2. Class conditional probability estimation

Recall the definition of class conditional probability $P_j$ ($j = 1, \ldots, k$) in § 3·1. In practice, after we build the classifier, it is important to estimate $P_j$ accordingly (Wang et al., 2008; Wu et al., 2010). In this section, we explore the relationship between the theoretical minimizer $f^*$ and the probability $P_j$. In practice, after obtaining the solution $\hat{f}$ to (1) or (2), one can replace $f^*$ by $\hat{f}$ to estimate $P_j$, as in the next theorem.

**Theorem 3—***In the angle-based classification structure, suppose the loss function $\ell$ is differentiable. For a given class conditional probability vector $(P_1, \ldots, P_k)$ and its corresponding theoretical minimizer $f^*$, the class probabilities can be expressed as*

$$P_j=\{\ell'(\langle W_j, f^*\rangle)^{-1}\}/\{\sum_{i=1}^{k}\ell'(\langle W_i, f^*\rangle)^{-1}\}(j=1,\ldots,k).$$

From Theorem 3, given the fitted $\hat{f}$, the estimated probabilities are $\hat{P}_j=\{\ell'(\langle W_j, \hat{f}\rangle)^{-1}\}/\{\sum_{i=1}^{k}\ell'(\langle W_i, \hat{f}\rangle)^{-1}\}(j=1,\ldots,k)$. If $\ell$ has a negative derivative $\ell'$, as in Theorem 1, one can verify that the estimated class conditional probabilities are proper, in the sense that $0 \le \hat{P}_j \le 1$ ($j = 1, \ldots, k$) and $\sum_{j=1}^{k}\hat{P}_j=1$.

In binary classification problems, it is known that the support vector machine does not provide class conditional probability estimation (Wang et al., 2008), because the hinge loss is not differentiable at 1. Liu et al. (2011) showed that when $c \to \infty$, $\mathrm{pr}(Y = 1 \mid X = x)$ is a step function, and thus cannot provide specific class conditional probability information.

In the multicategory case, probability estimation becomes more involved because we have a classification function vector instead of a single classification function. The probability estimation formula derived from Theorem 3 depends on $\hat{f}$ only through the derivative function $\ell'$. For $i \neq j$, $\hat{P}_i = \hat{P}_j$ if $\ell'(\langle W_i, \hat{f}\rangle) = \ell'(\langle W_j, \hat{f}\rangle)$. For the large-margin unified machine loss, $\ell'(u) = -1$ for $u \leq c/(1 + c)$. Thus, if $\hat{f}$ satisfies that $\langle W_i, \hat{f}\rangle \leq c/(1 + c)$ and $\langle W_j, \hat{f}\rangle \leq c/(1 + c)$, then the estimated class conditional probabilities for classes $i$ and $j$ are the same. As $c \to \infty$, $c/(1 + c) \to 1$, and the estimated probabilities are all $1/k$ for $\hat{f}$ with a norm small enough. For $i \neq j$, when $\langle W_j, \hat{f}\rangle$ is large and $\langle W_i, \hat{f}\rangle$ is small, we have $\hat{P}_j = 1$ and $\hat{P}_i = 0$. We show this phenomenon in Fig. 2 with $k = 3$. We see that in the multicategory angle-based large-margin unified machine case, as $c$ increases, the probability function becomes closer to a step function and thus the accuracy of probability estimation deteriorates. Our angle-based support vector machine defined in § 3·1 provides some probability information, while the angle-based classifier with the hinge loss does not, due to its nondifferentiability.

### 3·3. Asymptotic results

In this section, we extend the notations of excess $\ell$-risk and excess risk used by Bartlett et al. (2006) from the binary to the multicategory case, and study their convergence rates as the sample size $n \to \infty$. We focus on linear learning with a diverging number of covariates $p$, under the form (2). The penalty $J(f)$ we consider is the linear combination of the $L_1$ and $L_2$ penalties, similar to the elastic net penalty (Zou & Hastie, 2005). For simplicity, we assume that the intercepts are included in $J(f)$ by adding a variable to $x$ whose value is always 1. In this case, we can write $J(f)=\alpha\sum_{j=1}^{k-1}\|\beta_j\|_1+(1-\alpha)\sum_{j=1}^{k-1}\|\beta_j\|_2^2$, where $\beta_1, \ldots, \beta_{k-1}$ are vectors of length $p \geq 2$ that include the intercepts. Here $\|\cdot\|_1$ and $\|\cdot\|_2$ are the regular $L_1$ norm and $L_2$ norm of a vector, and the parameter $\alpha \in [0, 1]$ controls the relative proportion of the $L_1$ and $L_2$ penalties.

The $L_1$ penalty is well known for its variable selection property, and we show that the convergence rates of the excess $\ell$-risk for $\alpha > 0$ and $\alpha = 0$ are different. In particular, we show that when $\alpha > 0$, we have convergence provided $p = o\{\exp(n)\}$, and when $\alpha = 0$, the convergence requires that $p = o(n)$. Furthermore, we study the convergence rate of $\hat{\beta}_j$ to its best value, under the uniform metric. We then study the convergence rate of the excess risk using a conversion condition. Because we introduce some assumptions for the theory, we give an illustrative example in the Supplementary Material, where the assumptions are verified and the convergence rates are obtained.

For the remainder of § 3, we assume that the number of covariates including the intercept is $p = p_n$, and the penalty is $J(f) \leq s = s_n$, where both $p_n$ and $s_n$ may depend on $n$. We assume that each coordinate $x^{(l)} \in [0, 1]$ ($l = 1, \ldots, p_n$), and the loss function $\ell$ is Lipschitz with the Lipschitz constant 1. A function $\ell(\cdot)$ is said to be Lipschitz with constant $\gamma$ if for any $x_1$ and $x_2$ in its domain, $|\ell(x_1) - \ell(x_2)| \leq \gamma \|x_1 - x_2\|$, where $\|x_1 - x_2\|$ is some fixed norm in the domain of $\ell$. Our theory can be analogously extended to other bounded domains and Lipschitz functions with different Lipschitz constants. Because $p_n$ may become unbounded as $n \to \infty$, we assume that the underlying distribution $P(X, Y)$ is defined on $([0, 1]^\infty \times \{1,$

…, $k\}$, $\sigma^{\infty}([0,1]^{\infty}) \times 2^{\{1,\ldots,k\}}$) with $\sigma^{\infty}([0,1]^{\infty})$ being the σ-field generated by the open balls introduced by the uniform metric $d(x,x')=\sup_j|x_j-x_j'|$.

Before stating our theory, we first introduce some notation and definitions. For linear learning, we have that $f=(f_1,\ldots,f_{k-1}): f_j(x)=\beta_j^{\mathrm{T}}x (j=1,\ldots,k-1)$. The intercepts are included in the $\beta_j$ s, as discussed above. Let $F(p_n,s_n)=\{f=(f_1,\ldots,f_{k-1}): f_j(x)=\beta_j^{\mathrm{T}}x (j=1,\ldots,k-1), \alpha\sum_{j=1}^{k-1}\|\beta_j\|_1+(1-\alpha)\sum_{j=1}^{k-1}\|\beta_j\|_2^2 \le s_n\}$, and let $F(p_n) = \cup_{0\le s_n<\infty} F(p_n, s_n)$ be the full $p_n$-dimensional model. Suppose $\hat{f}=\mathrm{argmin}_{f\in F(p_n,s_n)} n^{-1}\sum_{i=1}^{n}\ell\{\langle W_{y_i}, f(x_i)\rangle\}$ is the empirical minimizer of the optimization problem (2), and $f^{(p_n)} = \mathrm{arginf}_{f\in F(p_n)} E[\ell\{\langle W_Y, f(x)\rangle\}]$ represents the best model under the full $p_n$-dimensional model. Here $f^{(p_n)}$ may not belong to $F(p_n)$.

Let $e_\ell(f, f^{(p_n)}) = E\{\ell(\langle W_Y, f\rangle)\} - E\{\ell(\langle W_Y, f^{(p_n)}\rangle)\}$. We call $e_\ell(f, f^{(p_n)})$ the excess $\ell$-risk, and name $e(f, f^{(p_n)}) = E\{L(Y, f)\} - E\{L(Y, f^{(p_n)})\}$ the excess risk, where $L$ is the 0–1 loss. Hence $E\{L(Y, f)\}$ is the generalization error of $f$. Our main result is as follows.

**Theorem 4—***Assume* $\tau_n = \{\log(p_n)/n\}1/2 \to 0$ *as* $n \to \infty$. *If* $\alpha > 0$, *then* $e_\ell(\hat{f}, f^{(p_n)})=O[\max\{s_n\tau_n\log(\tau_n^{-1}), d_n\}]$, *almost surely under P. If* $\alpha = 0$, *then* $e_\ell(\hat{f}, f^{(p_n)})=O[\max\{(p_n s_n)^{1/2}\tau_n\log(\tau_n^{-1}), d_n\}]$, *almost surely under P. Here* $d_n = \inf_{f\in F(p_n,s_n)} e_\ell(f, f^{(p_n)})$ *is the approximation error between* $F(p_n, s_n)$ *and* $F(p_n)$.

In Theorem 4, the balance between the approximation error and estimation error is represented by $d_n$ and $s_n$. In particular, as $s_n$ increases, the function class $F(p_n, s_n)$ becomes larger, the approximation error $d_n$ decreases and the estimation error increases. In this view, the best tradeoff takes place when $s_n\tau_n\log(\tau_n^{-1})\sim d_n$ for $\alpha > 0$, and $(p_n s_n)^{1/2}\tau_n\log(\tau_n^{-1})\sim d_n$ for $\alpha = 0$. If the model depends on finitely many predictors, we have a simpler version of Theorem 4.

**Assumption 1:** There exists a finite $s^*$, such that $f^{(p_n)} \in F(p_n, s^*)$ for all $p_n$.

Under Assumption 1, $d_n$ is strictly zero for large enough $n$ and $s_n$, leading to Corollary 1.

***Corollary 1:*** *When Assumption* 1 *is met, we have* $s_n = s^*$ *for all large n. Consequently, if* $\alpha > 0$, *then* $e_\ell(\hat{f}, f^{(p_n)})=O\{\tau_n\log(\tau_n^{-1})\}$, *almost surely under P. If* $\alpha = 0$, *then* $e_\ell(\hat{f}, f^{(p_n)})=O\{p_n^{1/2}\tau_n\log(\tau_n^{-1})\}$, *almost surely under P.*

There are important differences between the cases $\alpha > 0$ and $\alpha = 0$. For the former case, the convergence of $\tau_n\log(\tau_n^{-1})$ requires that $\tau_n = \{\log(p_n)/n\}^{1/2}$ converges, or in other words, $p_n$ grows no faster than $\exp(n)$. In the pure $L_2$ penalty case where $\alpha = 0$, the convergence of $p_n^{1/2}\tau_n\log(\tau_n^{-1})$ requires that $p_n = o(n)$, which is a much stronger assumption on the divergence speed of $p_n$. Theorem 4 and Corollary 1 shed some light on the effectiveness of the $L_1$ penalty when the true model contains many noise covariates.

We have established the convergence rate for the excess $\ell$-risk. As we focus on linear learning, the convergence rates for the estimated parameters $\hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$ are of interest as well. Next, we explore the convergence rate for $\hat{B} - B^{(p_n)}$, where $B = (\beta_1, \ldots, \beta_{k-1})$ is a $p_n$ by $k-1$ matrix whose columns are the parameters, $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_{k-1})$ is the estimated parameter matrix, and $B^{(p_n)} = (\beta_1^{(p_n)}, \ldots, \beta_{k-1}^{(p_n)})$ is the matrix whose columns represent the best parameters under the full $p_n$-dimensional model. In this section we assume that $p_n$ depends on $n$ and can go to infinity, hence we study the convergence rate with respect to the uniform metric, $d(B_1, B_2) = \sup_{i,j}\{(B_1)_{i,j} - (B_2)_{i,j}\}$, where $B_{i,j}$ is the $(i, j)$ element of the matrix $B$. First we introduce an assumption that is valid for many applications. For this assumption, the notation $X$ does not include the intercept term.

**Assumption 2:** The marginal distribution of $X$ is absolutely continuous with respect to the Lebesgue measure on $[0, 1]^{\infty}$.

Under Assumption 2, we study the convergence rate of $d(B, \hat{B}^{(p_n)})$ in the next theorem.

**Theorem 5—***Suppose $\ell$ is differentiable and convex, and Assumption* 2 *holds. We have* (a) *if the theoretical minimizer $f^* \in F(p_n)$ for some $p_n$, then $d(\hat{B}, B^{(p_n)}) = O\{\tau_n \log(\tau_n^{-1})\}^{1/2}$ for $\alpha > 0$ and $d(\hat{B}, B^{(p_n)}) = O\{p_n^{1/2}\tau_n \log(\tau_n^{-1})\}^{1/2}$ for $\alpha = 0$, almost surely under $P$,* (b) *if $f^* \notin F(\infty)$, and Assumption* 1 *holds, then $d(\hat{B}, B^{(p_n)}) = O\{\tau_n \log(\tau_n^{-1})\}$ for $\alpha > 0$ and $d(\hat{B}, B^{(p_n)}) = O\{p_n^{1/2}\tau_n \log(\tau_n^{-1})\}$ for $\alpha = 0$, almost surely under $P$.*

Theorem 5 indicates that the convergence rate of $\hat{B}$ to $B^{(p_n)}$ depends on whether the function class $F$ is large enough. In practice, the situation that $f^* \in F(p_n)$ for some $p_n$ rarely happens, and the convergence rate of $\hat{B}$ to $B^{(p_n)}$ is the same as the excess $\ell$-risk for most of the cases, if the true model is indeed sparse. We give an illustrative example in the Supplementary Material, where $f^* \in F(3)$, and the convergence rate of $\hat{B}$ to $B^{(p_n)}$ is slower than that of the excess $\ell$-risk.

We have studied the convergence rate of the excess $\ell$-risk. In practice, the convergence rate of excess risk is of interest if the classification accuracy of a given model converges to the best possible accuracy. In other words, we are interested in the consistency and convergence rate of the excess risk. Next, we establish a relationship between $e_\ell(f, \hat{f}^{(p_n)})$ and $|e(f, \hat{f}^{(p_n)})|$, where $e(f, \hat{f}^{(p_n)})$ is the difference between the misclassification rate of $f$ and $\hat{f}^{(p_n)}$. The technique of studying the excess risk using excess $\ell$-risk in the binary case was previously studied in Zhang (2004b) and Bartlett et al. (2006). Define the $L_2$ metric on $F$ as $\delta(f, f') = [\sum_{j=1}^{k-1} E\{f_j(X) - f_j'(X)\}^2]^{1/2}$. The following conversion assumption controls the behaviours of the excess $\ell$-risk and the excess risk in a small neighbourhood of $f^{(p_n)}$, introduced by the $L_2$ metric. It was previously used in Wang & Shen (2007) and Shen & Wang (2007).

**Assumption 3:** There exist constants $\gamma_1 \geq 1$ and $\gamma_2 > 0$ such that for all small $\varepsilon > 0$,

$$\inf_{\{f\in F(p_n):\delta(f,f^{(p_n)})\geq\varepsilon\}} e_\ell(f,f^{(p_n)}) \geq \xi_1(p_n)\varepsilon^{\gamma_1}, \quad (3)$$

$$\sup_{\{f\in F(p_n):\delta(f,f^{(p_n)})\leq\varepsilon\}} |e(f,f^{(p_n)})| \leq \xi_2(p_n)\varepsilon^{\gamma_2}, \quad (4)$$

where $\xi_1(p_n)$ and $\xi_2(p_n)$ may depend on $p_n$.

***Corollary 2:*** *Under Assumption 3, assume that* $\tau_n \to 0$ *as* $n \to \infty$. *Then*

$|e(\hat{f}, f^{(p_n)})| = O[\max\{s_n\tau_n\log(\tau_n^{-1}), d_n\}]^{\gamma_2/\gamma_1}$, *almost surely under P if* $\alpha > 0$, *and*

$|e(\hat{f}, f^{(p_n)})| = O[\max\{(p_n s_n)^{1/2}\tau_n\log(\tau_n^{-1}), d_n\}]^{\gamma_2/\gamma_1}$, *almost surely under P if* $\alpha = 0$. *Moreover, suppose Assumption 1 is also met, then if* $\alpha > 0$, *then*

$|e(\hat{f}, f^{(p_n)})| = O\{\tau_n\log(\tau_n^{-1})\}^{\gamma_2/\gamma_1}$, *almost surely under P. If* $\alpha = 0$, *then*

$|e(\hat{f}, f^{(p_n)})| = O\{p_n^{1/2}\tau_n\log(\tau_n^{-1})\}^{\gamma_2/\gamma_1}$, *almost surely under P.*

In the Supplementary Material, we give an example in which Assumptions 1, 2 and 3 are satisfied, and $\gamma_1$, $\gamma_2$ can be calculated explicitly.

### 3·4. Finite sample error bound

We use the surrogate loss $\ell$ as an upper bound of the 0–1 loss to make the optimization problem trackable, and the corresponding theoretical analysis is easier to deal with. We saw that Assumption 3 builds a relationship between the excess $\ell$-risk and the excess risk, and furthermore that one can establish the convergence rate of the excess risk by that of the excess $\ell$-risk. In this section, we study the finite sample bound on the expectation of the $\ell$-risk given $\hat{f}$, $E\{\ell(\langle W_Y, \hat{f}\rangle)\}$, which can be regarded as a bound on the misclassification rate of $\hat{f}$.

**Theorem 6—***The solution* $\hat{f}$ *to* (2) *satisfies that, with probability at least* $1 - \delta$,

$$E\{\ell(\langle W_Y, \hat{f}\rangle)\} \leq n^{-1}\sum_{i=1}^{n}\ell\{\langle W_{y_i}, \hat{f}(x_i)\rangle\} + Z,$$

*where*

$$Z = \begin{cases} s_n\alpha^{-1}A_1 + 4(k-1)s_n\alpha^{-1}n^{-1/4} + s_n\alpha^{-1}A_2, & \alpha > 0, \\ \{(k-1)p_n s_n\}^{1/2}A_1 + 4(k-1)(p_n s_n)^{1/2}n^{-1/4} + (p_n s_n)^{1/2}A_2, & \alpha = 0, \end{cases}$$

*with* $A_1 = 6\{\log(2/\delta)/n\}^{1/2}$ *and* $A_2 = 4n^{-1/4}[\log\{(e + 2ep_nk - 2ep_n)n^{-1/2}\}]^{1/2}$.

Theorem 6 gives an upper bound on $E\{\ell(\langle W_Y, \hat{f}\rangle)\}$ that depends only on $n$, $s_n$, $p_n$, $\alpha$ and the training sample, so it is directly computable from the data and the model we choose.

**Remark 1:** When solving the optimization (1), to calculate the result in Theorem 6, one may replace $s_n/\alpha$ by $\sum_{j=1}^{k-1}\|\hat{\beta}_j\|_1$ if $\alpha > 0$ and replace $p_n s_n$ by $\sum_{j=1}^{k-1}\|\hat{\beta}_j\|_2^2$ if $\alpha = 0$.

### 3·5. Comparison to existing methods with k classification functions

In this section, we provide some theoretical insight on the comparison between our angle-based method and regular multicategory large-margin classifiers using $k$ classification functions with the sum-to-zero constraint. We show that if the true signal in linear learning is sparse, then our angle-based method can enjoy a smaller estimation error, with the approximation error fixed at zero. Consequently, our method can give more accurate prediction.

We focus on comparing the complexity of functional classes of the corresponding optimization problems. To illustrate the idea, we use the method in Lee et al. (2004) as an example of the regular simultaneous methods. The proofs and conclusions are analogous for many other simultaneous classifiers. To begin with, we introduce some notation. Let $t(p_n, s_n) = s_n/\alpha$ if $\alpha > 0$ and $(p_n s_n)^{1/2}$ if $\alpha = 0$. Recall the definition of $F(p_n, s_n)$ in § 3·3. For our angle-based method, let $f^{(p_n,s_n)} = \text{argmin}_{f \in F(p_n,s_n)} E\{\ell(\langle f, W_Y\rangle)\}$. Define $h_f(\cdot) = \{2t(p_n, s_n)\}^{-1}\{\ell(\langle f, W_\cdot\rangle) - \ell(\langle f^{(p_n,s_n)}, W_\cdot\rangle)\}$, and let $H = \{h_f : f \in F(p_n, s_n)\}$. For the multicategory support vector machine of Lee et al. (2004), define

$$G(p_n, s_n)=\{f=(f_1,\ldots,f_k): f_j(x)=\beta_j^{\mathrm{T}}x (j=1,\ldots,k), \alpha\sum_{j=1}^{k}\|\beta_j\|_1+(1-\alpha)\sum_{j=1}^{k}\|\beta_j\|_2^2 \le s_n\}$$

with the sum-to-zero constraint. Let $f_L^{(p_n,s_n)}=\text{argmin}_{f\in G(p_n,s_n)} E\{\bar{L}(f,Y)\}$, where $\bar{L}$ is the multicategory support vector machine loss used by Lee et al. (2004). Analogously define $h_{L,f}(\cdot)=\{2t(p_n,s_n)\}^{-1}\{\bar{L}(f,\cdot)-\bar{L}(f_L^{(p_n,s_n)},\cdot)\}$, and $H_{\bar{L}} = \{h_{L,f} : f \in G(p_n, s_n)\}$. The next proposition provides the comparison between $H$ and $H_{\bar{L}}$ in terms of their uniform covering numbers $N(\varepsilon, \cdot)$. More details about uniform covering numbers are provided in the Supplementary Material.

**Proposition 1—***For positive* $\varepsilon$ *small enough,* $\log\{N(\varepsilon, H)\}$ *is bounded above by* $(2/\varepsilon^2)\log[e + e\{2p_n(k-1)\}\varepsilon^2] + \log(k-1)$, *and* $\log\{N(\varepsilon, H_{\bar{L}})\}$ *is bounded above by* $(2/\varepsilon^2)\log\{e + e(2p_n k)\varepsilon^2\} + \log(k)$.

From Proposition 1 and its proof, we can conclude that for fixed $\alpha$, the upper bound of $N(\varepsilon, H)$ of the angle-based methods is smaller than that of $N(\varepsilon, H_{\bar{L}})$ with $k$ functions, because our angle-based classifiers use only $k-1$ classification functions. In the Supplementary Material, we give an example where the upper bound of $N(\varepsilon, H_{\bar{L}})$ is almost tight. Furthermore, assume the true classification signal depends only on a finite number of predictors. In order to have the approximation error $d_n = 0$, one can verify that our angle-based method requires a smaller $s_n$ compared to the multicategory support vector machines in Lee et al. (2004). As a result, for a classification problem with sparse signal, our angle-based method can have a functional class smaller than that of Lee et al. (2004). Consequently, the angle-based method can have a smaller estimation error, which can lead to a better classification performance. See the proof of Theorem 4 in the Supplementary Material for more details on how the covering number affects the estimation error.

Intuitively, the regular simultaneous classifiers with $k$ functions can introduce extra variability in the estimated classifier, which can reduce the corresponding classification performance. Our angle-based method with $k - 1$ functions circumvents this difficulty and is more efficient.

As a remark, we point out that from Proposition 1, the difference of the uniform covering numbers becomes larger as the dimensionality $p_n$ increases. This suggests that the difference in classification performance between our angle-based methods and the regular multicategory large-margin classifiers can be large for high dimensional problems. We confirm this finding in § 5. In particular, we observe that for a classification problem with fixed and sparse signal, the difference in classification performance increases when the dimensionality increases.

## 4. Computational algorithms and tuning procedures

For a convex surrogate loss $\ell$ and a convex penalty term $J(f)$, (1) and (2) are convex problems and can be solved by many standard optimization tools (Boyd & Vandenberghe, 2004). For instance, with the squared loss $\ell(u)=(1-u)^2$ and the $L_2$ penalty, there exist explicit solutions for the problem (1), or one can obtain its equivalent solution by solving the Karush–Kuhn–Tucker conditions generated from (2). If one applies the generalized hinge loss with a reject option (Bartlett & Wegkamp, 2008) and the $L_1$ penalty, then linear programming can be used to solve (2). In this section, we demonstrate how to implement (1) using the large-margin unified machine loss with finite $c$ and the $L_2$ penalty, by the coordinate descent algorithm (Friedman et al., 2010).

For simplicity we assume the intercepts are included in the parameters, as in § 3·3. Let the estimated parameters at the $m$th step be $\hat{B}^{(m)}=(\hat{\beta}_1^{(m)},\ldots,\hat{\beta}_{k-1}^{(m)})$. For clarification, the notation $B^{(\hat{m})}$ is different from $B^{(p_n)}$ defined in § 3·3. At the $(m+1)$th step, let $W_{i,j}(z)=\sum_{q<l}x_{iq}\hat{\beta}_j^{(m+1),(q)}+\sum_{q>l}x_{iq}\hat{\beta}_j^{(m),(q)}+zx_{i,l}$, where $\hat{\beta}_j^{(m+1),(q)}$ is the $q$th element of $\hat{\beta}_j^{(m+1)}$ and $\hat{\beta}_j^{(m),(q)}$ is defined in a similar way. Then $W_{i,j}(z)$ is a function of $z\in\mathbb{R}$. Furthermore, define

$$\tilde{f}_{i,j,-l}(z)=\left(x_i^{\mathrm{T}}\hat{\beta}_1^{(m+1)},\ldots,x_i^{\mathrm{T}}\hat{\beta}_{j-1}^{(m+1)},W_{i,j}(z),x_i^{\mathrm{T}}\hat{\beta}_{j+1}^{(m)},\ldots,x_i^{\mathrm{T}}\hat{\beta}_{k-1}^{(m)}\right)^{\mathrm{T}}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, k-1$ and $l = 1, \ldots, p$. Set

$$\hat{\beta}_j^{(m+1),(l)}=\operatorname*{argmin}_z\left[\lambda z^2+n^{-1}\sum_{i=1}^{n}\ell\{\langle\tilde{f}_{i,j,-l}(z),W_{y_i}\rangle\}\right]. \quad (5)$$

The optimization (5) is a one-dimensional problem, and so can be solved efficiently. We update all the elements in $\hat{\beta}_j^{(m+1)}(j=1,\ldots,k-1)$ until convergence.

We summarize the above coordinate descent algorithm in Algorithm 1.

**Algorithm 1**

| | |
|---|---|
| *Step* 1. | Initialize the algorithm with $\hat{\beta_j} = (0, \ldots, 0)^T$, for all $j = 1, \ldots, k-1$. |
| *Step* 2. | For the $m$th loop, with the solution $(\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_{k-1}^{(m)})$ given, update the elements of $\hat{\beta_1}$ sequentially. After $\hat{\beta_1}$ has been updated, update $\hat{\beta_2}, \ldots, \hat{\beta_{k-1}}$ in the same manner. |
| *Step* 3. | Repeat Step 2 until convergence. |

For the regular classifiers with the sum-to-zero constraint, one can remove the constraint and reparameterize one function as the negative sum of others. This can be computationally inefficient compared to the angle-based method. We take the coordinate descent algorithm as an example. Without loss of generality we assume that $f_k$ is reparameterized. To update a parameter in $f_1$, one needs to involve all the terms that depend on $f_1$ and $f_k$ to calculate the next updated value. This can be much slower than the angle-based method, which requires only the terms that depend on $f_1$. For other optimization tools, such as the linear or quadratic programming for the simultaneous support vector machines (Lee et al., 2004;Wang & Shen, 2007; Liu & Yuan, 2011), reparameterizing can be even slower than having the sum-to-zero constraint in the optimization. Therefore, our angle-based method can be much faster.

In practice, a proper choice of the tuning parameter $\lambda$ or $s$ is crucial for the accuracy of our classifier. Different tuning parameters correspond to different prediction models. There are various tuning techniques in the literature. Here, we briefly discuss the crossvalidation procedure, which is commonly used in practice. For the crossvalidation approach, one partitions the training dataset into $q$ subsets of observations whose sizes are roughly the same. Each time a single subset of observations is used for tuning, and the remaining $q - 1$ subsets are used for training. One repeats the process $q$ times when all observations have been used for tuning, and the $q$ prediction results are combined to select the best tuning parameter.

## 5. Simulation examples

In this section, we use five simulated examples to explore the performance of our angle-based method. For each example, we apply penalized logistic regression, the support vector machine approximated by a large-margin unified machine with large $c$, and the tuned large-margin unified machine (Liu et al., 2011). We also implement other existing methods with the sum-to-zero constraint, including several multicategory support vector machine formulations (Vapnik, 1998; Crammer & Singer, 2001; Lee et al., 2004) and the penalized logistic regression proposed by Zhu & Hastie (2005). We show that our angle-based classifiers can give more accurate classification, and their computational cost can be significantly smaller than these alternative methods.

In the first four simulated examples, we generate datasets whose signal depends linearly on a few covariates, and we add additional covariates that are pure noise variables. The total dimensions are set to be 10, 50, 100 and 200, and we observe the pattern of change in classification performance. For the fifth example, we perform Gaussian kernel learning.

Instead of letting the dimensionality grow, we fix the dimension of this example and let the number of observations increase.

For classification performance, we compare the test error rates and probability estimation using the mean absolute error, $E(|p - \hat{p}|)$, on a test set of size $10^5$. By Theorem 3, our angle-based approximate support vector machine can provide class conditional probability estimation. Within one replication, the model is built on a training dataset, and the optimal tuning parameter is chosen by minimizing the classification error rate over an independent tuning set with a grid search. For all the classifiers including our angle-based methods and the existing ones, we choose the $L_2$ penalty as the regularization term $J(f)$. Through 1000 replicates, we record the total computational time for training a model, tuning it on 30 different values of tuning parameters, and making prediction on the test set. We report the average time in seconds as a measurement of speed. All simulations are done using R (R Development Core Team, 2014), on a 2·80 GHz Intel processor. We report the simulation setting and results of Examples 2 to 5 in the Supplementary Material.

### Example 1

We generate a three-class dataset, where $\text{pr}(Y = j) = 1/3$, and $(X \mid Y = j) \sim N(\mu_j, \sigma^2 I_2)$ ($j = 1, 2, 3$). Here $\mu_j$s are equally distributed on the unit circle, and $\sigma$ is chosen such that the Bayes error is 0·1. The noise covariates are independent and identically distributed as $N(0, 0{\cdot}5)$. Both the training and tuning datasets are of size 100.

Table 1 reports the behaviours of the classifiers. The angle-based classifiers have smaller error rates overall, while the tuned angle-based large-margin unified machine works best. For the four linear learning examples, the difference in classification accuracy between angle-based methods and the regular classifiers is small when the dimension is low. When the dimension gets higher, the difference becomes larger. This is consistent with the theoretical insights provided in § 3·5. For the kernel learning example, the angle-based classifiers are also very competitive. For all the examples, the computational costs of the angle-based methods are significantly less than those of the other methods. In terms of probability estimation, the approaches of Vapnik (1998), Crammer & Singer (2001) and Lee et al. (2004) do not provide class conditional probability information. In contrast, our angle-based methods can estimate class conditional probabilities, and yield more accurate results than the penalized logistic classifier.

## 6. Real data examples

In this section, we demonstrate the performance of our angle-based classifiers via three datasets, CNAE-9, Semeion Handwritten Digit and Vehicle from the University of California Irvine machine learning repository website, and a recent Glioblastoma Multiforme Cancer gene expression dataset. See the Supplementary Material for more information on the Glioblastoma Multiforme Cancer data. To select the best tuning parameter, we split each dataset into six groups of observations whose sizes are roughly the same, choose one group as the testing data, and perform 5-fold crossvalidations on the remaining observations. We perform linear learning on the CNAE-9, Semeion Handwritten Digit and Glioblastoma data, and we use second-order polynomial kernel learning for the

Vehicle data. To reduce the computational cost in the Glioblastoma dataset, we choose 1000 genes with the largest median absolute deviation values based on the training sample within each replication.

The results are reported in Table 2. The classification accuracy of our angle-based methods is better than that of the other methods, and the angle-based tuned large-margin unified machine loss works the best overall. Furthermore, the computational time for the angle-based methods is considerably shorter, consistent with the simulation results.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgement

The authors thank the editor, the associate editor and two reviewers for their helpful suggestions. This work was supported in part by the U.S. National Institutes of Health and National Science Foundation. Yufeng Liu is also affiliated with the Carolina Center for Genome Sciences and the Department of Biostatistics at the University of North Carolina.

## References

1. Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. J. Am. Statist. Assoc. 2006; 101:138–156.
2. Bartlett PL, Wegkamp MH. Classification with a reject option using a hinge loss. J. Mach. Learn. Res. 2008; 9:1823–1840.
3. Boser, BE.; Guyon, IM.; Vapnik, VN. A training algorithm for optimal margin classifiers. In: Haussler, D., editor. Proc. 5th Ann. Workshop Comp. Learn. Theory. New York: Association for Computing Machinery; 1992. p. 144-152.COLT ’92.
4. Boyd, SP.; Vandenberghe, L. Convex Optimization. Cambridge: Cambridge University Press; 2004.
5. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. 2001; 2:265–292.
6. Cristianini, N.; Shawe-Taylor, JS. An Introduction to Support Vector Machines. Cambridge: Cambridge University Press; 2000.
7. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comp. Syst. Sci. 1997; 55:119–139.
8. Friedman JH, Hastie TJ, Tibshirani RJ. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software. 2010; 33:1–22.
9. Hastie, TJ.; Tibshirani, RJ.; Friedman, JH. The Elements of Statistical Learning. 2nd ed.. New York: Springer; 2009.
10. Hill SI, Doucet A. A framework for kernel-based multi-category classification. J. Artif. Intel. Res. 2007; 30:525–564.
11. Lange K, Wu TT. An MM algorithm for multicategory vertex discriminant analysis. J. Comp. Graph. Statist. 2008; 17:527–544.
12. Lee Y, Lin Y, Wahba G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. J. Am. Statist. Assoc. 2004; 99:67–81.
13. Lin X, Wahba G, Xiang D, Gao F, Klein R, Klein B. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. Ann. Statist. 2000; 28:1570–1600.
14. Liu Y, Shen X. Multicategory ψ-learning. J. Am. Statist. Assoc. 2006; 101:500–509.
15. Liu Y, Yuan M. Reinforced multicategory support vector machines. J. Comp. Graph. Statist. 2011; 20:901–919.

16. Liu Y, Zhang HH, Wu Y. Soft or hard classification? Large margin unified machines. J. Am. Statist. Assoc. 2011; 106:166–177.
17. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. ISBN 3-900051-07-0. http://www.R-project.org.
18. Saberian, MJ.; Vasconcelos, N. Multiclass boosting: Theory and algorithms. In: Shawe-Taylor, JS.; Zemel, RS.; Bartlett, PL.; Pereira, FCN.; Weinberger, KQ., editors. Adv. Neural Info. Proces. Syst. Vol. 24. 2011. p. 2124-2132.
19. Shen X, Tseng GC, Zhang X, Wong WH. On ψ-learning. J. Am. Statist. Assoc. 2003; 98:724–734.
20. Shen X, Wang L. Generalization error for multi-class margin classification. Electron. J. Statist. 2007; 1:307–330.
21. Tewari A, Bartlett PL. On the consistency of multiclass classification methods. J. Mach. Learn. Res. 2007; 8:1007–1025.
22. Vapnik, VN. Statistical Learning Theory. New York: Wiley; 1998.
23. Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. Adv. Kernel Meth. Support Vector Learn. 1999; 6:69–87.
24. Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. Nat. Acad. Sci. 2002; 99:16524–16530. [PubMed: 12477931]
25. Wang J, Shen X, Liu Y. Probability estimation for large margin classifiers. Biometrika. 2008; 95:149–167.
26. Wang L, Shen X. On L1-norm multi-class support vector machines: Methodology and theory. J. Am. Statist. Assoc. 2007; 102:595–602.
27. Wu TT, Wu Y. Nonlinear vertex discriminant analysis with reproducing kernels. Statist. Anal. Data Mining. 2012; 5:167–176.
28. Wu Y, Zhang HH, Liu Y. Robust model-free multiclass probability estimation. J. Am. Statist. Assoc. 2010; 105:424–436.
29. Zhang C, Liu Y. Multicategory large-margin unified machines. J. Mach. Learn. Res. 2013; 14:1349–1386. [PubMed: 24415909]
30. Zhang T. Statistical analysis of some multi-category large margin classification methods. J. Mach. Learn. Res. 2004a; 5:1225–1251.
31. Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Statist. 2004b; 32:56–85.
32. Zhu J, Hastie TJ. Kernel logistic regression and the import vector machine. J. Comp. Graph. Statist. 2005; 14:185–205.
33. Zhu J, Zou H, Rosset S, Hastie TJ. Multi-class Adaboost. Statist. Interf. 2009; 2:349–360.
34. Zou H, Hastie TJ. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B. 2005; 67:301–320.

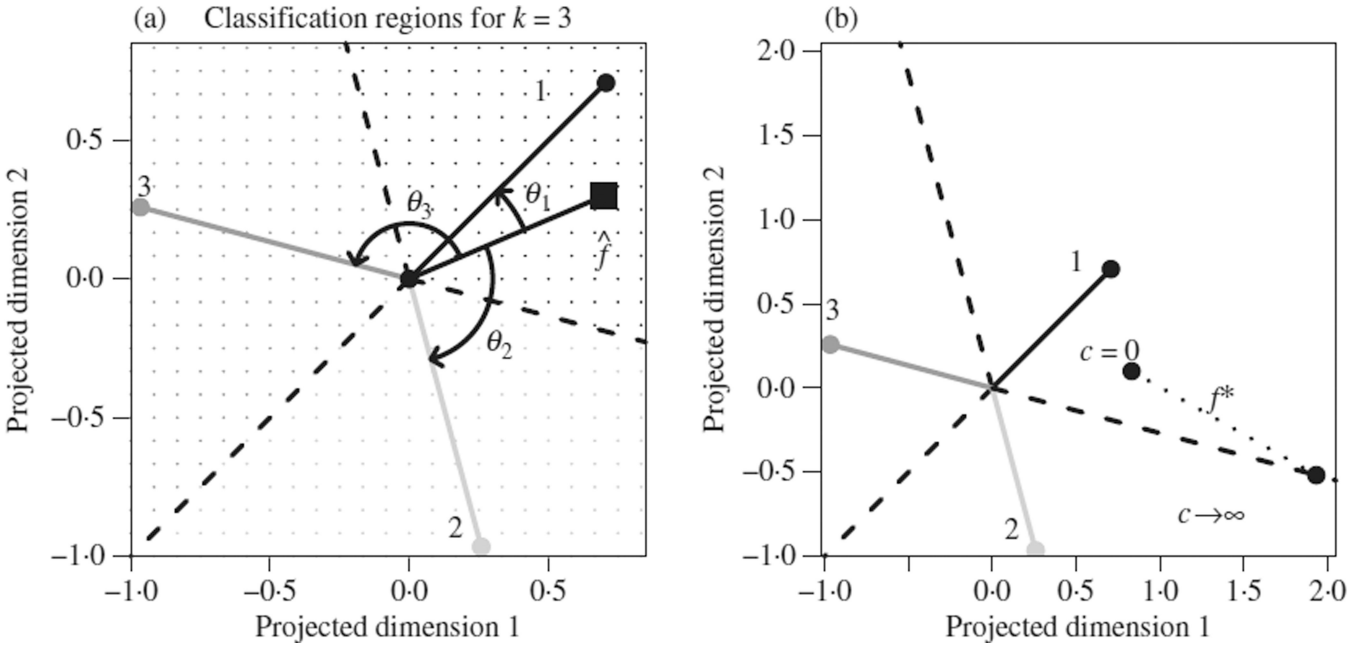


**Fig. 1.**

Illustration for the angle-based classification with $k = 3$. (a) The classification regions for $k = 3$. The mapped observation $\hat{f}$ is predicted as class 1, because $\theta_1 < \theta_2 < \theta_3$, where $\theta_j$ ($j = 1, 2, 3$) are shown in the figure. Observe that $\mathbb{R}^2$ is naturally divided into three classification regions, together with the classification boundaries. (b) The plot of $f^*$ with $P = (0{\cdot}55, 0{\cdot}25, 0{\cdot}2)$ and $a = 1$. Observe that $f^*$ moves along the dotted line as $c$ changes from 0 to $\infty$. For $c < \infty$, $f^*$ remains in $C_1$, and hence the angle-based large-margin unified machine is Fisher consistent. When $c \to \infty$, $f^*$ is on the boundary, and consequently the angle-based support vector machine with hinge loss is not Fisher consistent.

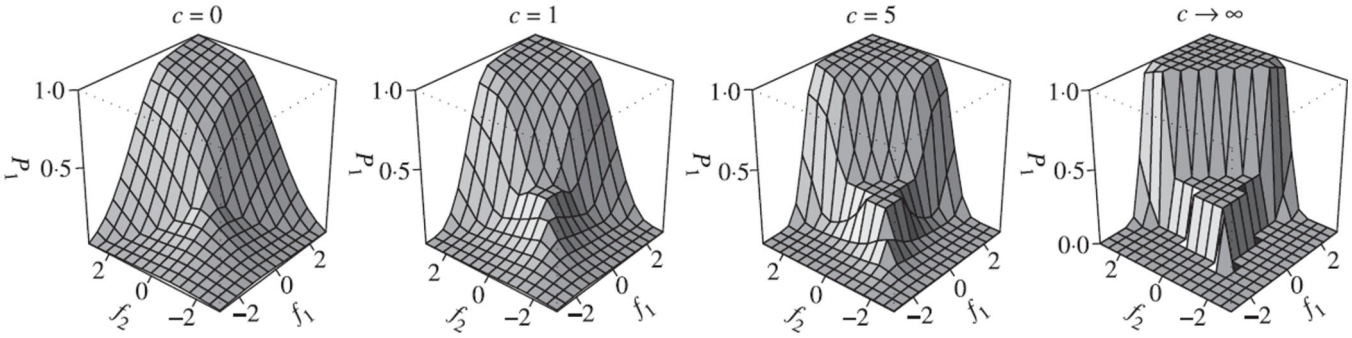


**Fig. 2.**

Visualization of the relationship between $f_1^*, f_2^*$ and $P_1$, with $k = 3$, $a = 1$, and $c \in \{0, 1, 5\}$ and $c \to \infty$. As $c$ increases, the function becomes closer to a step function, with more probability information lost.

## Algorithm 1

| | |
|---|---|
| *Step* 1. | Initialize the algorithm with $\hat{\beta_j} = (0, \ldots, 0)^{\mathrm{T}}$, for all $j = 1, \ldots, k - 1$. |
| *Step* 2. | For the $m$th loop, with the solution $(\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_{k-1}^{(m)})$ given, update the elements of $\hat{\beta_1}$ sequentially. After $\hat{\beta_1}$ has been updated, update $\hat{\beta_2}, \ldots, \hat{\beta_{k-1}}$ in the same manner. |
| *Step* 3. | Repeat Step 2 until convergence. |

**Table 1**

Summary of the classification results for Example 1

| | | Dimension 10 | | | Dimension 50 | | | Dimension 100 | | | Dimension 200 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Example 1** | | **Error** | **Prob** | **Time** | **Error** | **Prob** | **Time** | **Error** | **Prob** | **Time** | **Error** | **Prob** | **Time** |
| Existing methods | SVM1 | 12·5 | NA | 14 | 24·9 | NA | 20 | 34·4 | NA | 27 | 43·3 | NA | 37 |
| | SVM2 | 12·3 | NA | 12 | 24·6 | NA | 21 | 34·2 | NA | 28 | 43·3 | NA | 37 |
| | SVM3 | 12·6 | NA | 13 | 25·1 | NA | 24 | 34·4 | NA | 27 | 42·7 | NA | 41 |
| | Logi | 12·0 | 9·2 | 16 | 24·4 | 16·2 | 21 | 34·0 | 17·7 | 30 | 41·2 | 19·5 | 39 |
| Angle based | ALogi | 11·2 | 7·3 | 7 | 17·8 | 10·1 | 9 | 23·0 | 12·0 | 12 | 33·3 | 12·9 | 17 |
| | ASVM | 11·9 | 12·9 | 9 | 19·0 | 13·7 | 11 | 24·6 | 15·5 | 16 | 34·5 | 16·2 | 22 |
| | ALUM | 11·3 | 7·3 | 15 | 17·7 | 11·2 | 19 | 22·9 | 11·7 | 27 | 33·1 | 12·9 | 38 |

The standard deviations of the mean error rates range from 0·04 to 0·23. SVM1, support vector machines of Vapnik (1998); SVM2, support vector machines of Crammer&Singer (2001); SVM3, support vector machines of Lee et al. (2004); Logi, multicategory logistic regression of Zhu & Hastie (2005); ALogi, angle-based logistic regression; ASVM, angle-based support vector machine using approximated hinge loss; ALUM, angle-based tuned large-margin unified machines of Liu et al. (2011); Error, classification error percentage; Prob, the mean percentage of absolute error on class conditional probability estimation.

**Table 2**

Summary of the classification results for the real datasets

| | | CNAE-9 | | Semeion | | Vehicle | | Glioblastoma | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Error** | **Time** | **Error** | **Time** | **Error** | **Time** | **Error** | **Time** |
| Existing methods | SVM1 | 14·1 | 241 | 14·0 | 110 | 25·4 | 122 | 20·3 | 877 |
| | SVM2 | 13·6 | 266 | 14·2 | 120 | 24·9 | 116 | 20·1 | 910 |
| | SVM3 | 13·1 | 241 | 14·6 | 131 | 26·0 | 142 | 19·7 | 883 |
| | Logi | 15·7 | 203 | 15·1 | 105 | 26·7 | 109 | 20·0 | 677 |
| Angle based | ALogi | 12·5 | 101 | 12·9 | 51 | 23·9 | 44 | 18·1 | 359 |
| | ASVM | 12·2 | 138 | 12·6 | 69 | 23·3 | 67 | 17·9 | 497 |
| | ALUM | 12·2 | 199 | 12·5 | 103 | 23·5 | 104 | 17·7 | 810 |

The standard deviations of the mean error rates range from 0·04 to 0·15. See Table 1 for abbreviations.